Exponential Reduction in Sample Complexity with Learning of Ising Model Dynamics

Arkopal Dutt [1] **Andrey Y. Lokhov** [2] **Marc Vuffray** [2] **Sidhant Misra** [2]

Abstract

The usual setting for learning the structure and parameters of a graphical model assumes the availability of independent samples produced from the corresponding multivariate probability distribution. However, for many models the mixing time of the respective Markov chain can be very large and i.i.d. samples may not be obtained. We study the problem of reconstructing binary graphical models from correlated samples produced by a dynamical process, which is natural in many applications. We analyze the sample complexity of two estimators that are based on the interaction screening objective and the conditional likelihood loss. We observe that for samples coming from a dynamical process far from equilibrium, the sample complexity reduces exponentially compared to a dynamical process that mixes quickly.

1. Introduction

A graphical model (GM) is a convenient description of a probabilistic distribution which highlights the structure of the conditional dependencies existing between a set of random variables. We focus our attention on GMs that can be expressed as elements of an exponential family and are naturally associated with a graph that captures the underlying structure of the conditional dependencies. These GMs, sometimes referred as positive Markov random fields or Boltzmann distributions, are ubiquitous tools used to describe behaviors of random systems across a broad range of sciences such as physics (Chaves et al., 2015), biology (Jansen et al.), medicine (Constantinou et al., 2016), data mining (Buczak & Guven, 2016) and computer vision (Wang et al., 2013). The expression of GMs can sometimes be deduced from first principles, but often it has to be learned from observed data accessible through measurements and experiments. As these samples are time-consuming or costly to produce, it is not surprising that *efficient* GM learning methods play an important role in various fields such as in the study of gene expression (Marbach et al., 2012), protein interactions (Morcos et al., 2011), neuroscience (Schneidman et al., 2006a), image processing (Roth & Black, 2005), sociology (Eagle et al., 2009) and even grid science (He & Zhang, 2011).

The practical problem of learning a GM from observed data has a long-standing and rich history that can be traced back to the seminal work of Chow-Liu (Chow & Liu, 1968). However, it wasn't until recently and after further developments that a body of work showed one can efficiently reconstruct GMs from independent and identically distributed (i.i.d.) samples (Bresler, 2015; Vuffray et al., 2016; Hamilton et al., 2017; Klivans & Meka, 2017; Lokhov et al., 2018; Vuffray et al., 2019). In these papers, two methods stand out for being essentially optimal in the number of samples that they require (Lokhov et al., 2018). These methods are named Regularized Interaction Screening Estimator (RISE) and Regularized Pseudo-likelihood Estimator (RPLE) and both rely on the minimization of a convex loss function. The sample complexity of these estimators scales exponentially with a quantity named β that represents the maximum magnitude of the parameters in the GM. This exponential dependence in β is a fundamental limit of GM learning from i.i.d. samples (Santhanam & Wainwright, 2012) with heavy practical consequences as it restricts the possibility of learning GMs when data is scarce. However, the assumption of having access to independent samples is a modeling hypothesis that is convenient in many ways, but for which we can challenge the limits of its validity as it is known that sampling from arbitrary GMs is an NP-hard task. In most of the experimental settings mentioned earlier, the samples are actually obtained from a dynamic process whose stationary distribution is captured by a GM. Even the state of the art sampling techniques for GMs are implemented through Markov Chain Monte-Carlo (MCMC) dynamics (Levin & Peres, 2017; Gotovos et al., 2015). It is therefore natural to wonder if learning a graphical model from a dynamical process can be beneficial from a sample complexity standpoint.

Surprisingly, GM learning from dynamics has been rigor-

[1]Massachusetts Institute of Technology, Cambridge, MA, USA [2]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, USA. Correspondence to: Andrey Y. Lokhov <lokhov@lanl.gov>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

ously studied very little with the notable exception of the paper of Bresler, Gamarnik and Shah (Bresler et al., 2017). In an attempt to demonstrate that learning GMs from non-i.i.d. samples can be tractable, a question that was still widely debated at the time of the paper's initial release, Bresler, Gamarnik and Shah proved that one can efficiently learn GMs using samples coming from Glauber dynamics (Glauber, 1963), an iconic MCMC sampling dynamics. This result was regrettably overshadowed by the progress made in the following years in GM learning and their algorithm suffers from an impractical scaling much worse compared to what one could obtain with RISE or RPLE in the i.i.d. sample setting. The question of whether correlated samples from dynamics can improve the sample complexity of GM learning remains unanswered.

A hint on the fact that such a reduction in sample complexity is possible is provided by a number of empirical studies in the statistical physics literature that considered reconstruction using mean-field methods (Roudi & Hertz, 2011; Mézard & Sakellariou, 2011; Zeng et al., 2011; Zhang, 2012; Bachschmid-Romano & Opper, 2015) and using pseudo-likelihood (Besag, 1975) based estimators (Zeng et al., 2013; Decelle & Zhang, 2015; Decelle et al., 2016) in various settings, although most of these studies focus on a simpler setting of asymmetric couplings known as kinetic Ising model that does not contain the GM as its equilibrium state. We do not consider mean-field based methods here because these methods are not exact and typically only work for high-temperature (weakly coupled) models, see (Lokhov et al., 2018) for an extensive discussion on the value of exact algorithms. Existing studies of pseudo-likelihood based estimators have been mostly conducted in a setting of reconstruction of single instances, and with a focus on parameter estimation (instead of structure learning for which the sample complexity bounds are known); hence, it is hard to extract the sample complexity scalings with model parameters such as β from these works. Still, single-instance reconstruction results indicate that in practice the number of samples required for an accurate model learning in the dynamic case seems to be significantly smaller compared to the i.i.d. learning setting.

In this work, we quantify through a carefully designed set of experiments and a rigorous mathematical analysis the reduction in sample complexity that one can achieve using samples from Glauber dynamics. We focus our attention on Ising models, the celebrated class of pairwise and binary GMs for which information-theoretic lower-bounds on sample complexity exist both for i.i.d. samples (Santhanam & Wainwright, 2012) and samples coming from Glauber dynamics (Bresler et al., 2017). We propose an adaptation of the efficient learning algorithms RISE and RPLE for learning GMs with dynamical samples; Interaction Screening method has never been previously considered for learning in the dynamic setting. We extract the β scaling of the sample complexity for different instances of Ising models in two different dynamical regimes. The first, denoted as T-regime, consists in learning an Ising model from a single Glauber dynamic trajectory that mixes quickly toward its stationary distribution. The second, referred to as M-regime, consists in learning an Ising model from a series of one step evolutions of the Glauber dynamics from an initial distribution thus mimicking the trajectory of a system far from its mixed state. A similar setting of learning from a number of short trajectories starting with uniformly sampled configurations instead of one long trajectory has been considered in (Decelle et al., 2016). We find that the β scaling in the T-regime is similar to the one obtained from learning GMs with i.i.d. samples, an expected result since the Glauber dynamics produces i.i.d. samples once it has mixed. However, our main finding is that in the M-regime the β scaling depends crucially on the initial distribution, and for dynamics far from equilibrium we achieve a β exponent scaling up to ten times better than in the i.i.d. case. This exponential improvement in the sample complexity concretely translates into a reduction in sample requirements by a factor $10^4 - 10^5$ in typical regimes where variables of the GMs display non-trivial correlations. Our results also have a deep theoretical implication as we show that samples acquired far from the equilibrium carry more information about the structure of the problem. Based on this intuition, we design an active learning algorithm that modifies the trajectory of the dynamics on the fly to optimize the sample complexity of the learning task.

The paper is organized as follows. In Sec. 2, we define the problem of learning an Ising model from Glauber dynamics and describe two different regimes under which learning can take place. In Sec. 3, we present our learning algorithms and a theoretical analysis of their scaling properties. Additionally, we assess their performance experimentally on a variety of Ising models of different topologies and interaction strengths. In Sec. 4, we illustrate a real world application of our algorithms and present how active learning can be used to gain further advantage in learning from dynamics. The conclusion can be found in Sec. 5.

2. Problem statement

2.1. Ising model

Consider the Ising model on a graph $G = (V, E)$ with n nodes where $V = [n]$ is the set of nodes and $E \subset V \times V$ is the set of undirected edges. Each node $i \in V$ is associated with a spin which we will denote by σ_i and is a binary random variable taking values in $\{-1, +1\}$. The neighborhood of a node i is denoted by $\partial i = \{j \in V \mid (i, j) \in E\}$. The probability measure of a particular configuration of spins $\underline{\sigma} \in \{-1, +1\}^n$ is given by the Gibbs distribution

$$p(\underline{\sigma}) = \frac{1}{Z} \exp\left(\sum_{(i,j)\in E} J^\star_{ij}\sigma_i\sigma_j + \sum_{i\in V} H^\star_i \sigma_i \right), \quad (1)$$

where $\underline{J}^\star = \{J^\star_{ij}\}_{(i,j)\in E}$ is the vector of non-zero interactions associated with each edge, and $\underline{H}^\star = \{H^\star_i\}_{i\in V}$ is the vector of magnetic fields associated with each node. The normalization factor $Z = \sum_{\underline{\sigma}} \exp\left(\sum_{(i,j)\in E} J^\star_{ij}\sigma_i\sigma_j + \sum_{i\in V} H^\star_i\sigma_i\right)$ is referred to as the partition function and is in general NP-hard to compute (Sly & Sun, 2012).

2.2. Glauber dynamics and observations

Glauber dynamics is a reversible Markov chain that was originally introduced in (Glauber, 1963) for Ising models and can be generalized for any Markov random field. The Glauber dynamics is specified by the update rule that determines its transition probabilities. The spin configuration at any time t is denoted by $\underline{\sigma}^t$ with the initial configuration being $\underline{\sigma}^0$. At each time step t, a node is chosen uniformly at random. The corresponding random variable is given by I^{t+1}. Conditioned on $I^{t+1} = i$, the spin σ_i is updated according to the following conditional distribution:

$$p(\sigma_i^{t+1}|\underline{\sigma}^t) = \frac{\exp\left[\sigma_i^{t+1}(\sum_{j\in\partial i} J^\star_{ij}\sigma^t_j + H^\star_i)\right]}{2\cosh\left[\sum_{j\in\partial i} J^\star_{ij}\sigma^t_j + H^\star_i\right]}. \quad (2)$$

The initial configuration $\underline{\sigma}^0$ is assumed to be drawn from some distribution $p_0(\underline{\sigma}^0)$. Executing m steps of the Glauber dynamics yields the samples $\underline{\sigma}^1, \underline{\sigma}^2, ..., \underline{\sigma}^m$ and the corresponding sequence of node identities is then $I^1, I^2, ..., I^m$. It can be used to draw i.i.d. samples from the Gibbs distribution in (1) when run long enough to allow for mixing. However, for a large class of models this mixing time is exponentially high (Martinelli & Olivieri, 1994), limiting its computational tractability. At the same time, many out-of-equilibrium natural systems such as biological neural networks naturally generate temporally correlated spike train data (Berry et al., 1997; Pillow et al., 2008) that is well described and is modeled by the Glauber dynamics (Marre et al., 2009; Tyrcha et al., 2013). This raises the problem of learning the graphical model associated with the sequence of time-correlated samples produced by Glauber dynamics, with the goals of inferring the connectivity of the system, predicting the final state of the dynamics, or for building a reliable model that can be used to simulate and predict the dynamics starting from other configurations.

2.3. Glauber dynamics with multi-start and the model selection problem

Suppose that the Glauber dynamics is run in batches of size m_r for $r = 1, \ldots, R$ with total number of samples $\sum_r m_r = m$. For each r, the initial configuration is picked according to the probability distribution p_0 and a sequence of m_r steps of the Glauber dynamics are executed to obtain m_r samples. In this paper, we consider two extreme cases. The **T-regime** corresponds to $R = 1$ and $m_1 = m$ where starting from an initial configuration, one batch of size m is executed to obtain m samples. The **M-regime** corresponds to $m_r = 1\ \forall r$ and executing one step of the Glauber dynamics m times, each time starting from a new initial configuration. The two regimes are designed to emulate the behaviour of the Markov chain close to the equilibrium distribution (T-regime), and far from the equilibrium distribution (M-regime).

Notation: We will denote a sample by a tuple of the input spin configuration to a step of Glauber dynamics, the resulting spin configuration of Glauber dynamics and the updated node identity. The sample produced in the $(t+1)$-th step of T-regime is given by $(\underline{\sigma}^t, \underline{\sigma}^{t+1}, I^{t+1})$ and the t-th sample produced in M-regime will be given by $\{\underline{\sigma}^{0(t)}, \underline{\sigma}^{1(t)}, I^{1(t)})$. Note the difference in the superscripts. For convenience, we will denote the set $\{1, 2, ..., k\} = [k]$ for $k \in \mathbb{Z}^+$. For stating sample complexity results, it is convenient to define a minimum non-zero coupling $\alpha = \min_{(i,j)\in E}|J^\star_{ij}|$, the maximum coupling strength $\beta = \max_{(i,j)\in E}|J^\star_{ij}|$, and a maximum nodal degree d for the Ising model in (1).

The dynamic model selection problem: Given m samples of $\{(\underline{\sigma}^t, \underline{\sigma}^{t+1}, I^{t+1})\}_{t=0,\ldots,m-1}$ or $\{\underline{\sigma}^{0(t)}, \underline{\sigma}^{1(t)}, I^{1(t)}\}_{t\in[m]}$ observed from the Glauber dynamics from either the T-regime or M-regime, the model selection problem consists of two parts:

1. *Parameter estimation:* Compute estimates $\hat{J}_{ij}$ of $J^\star_{ij}$ such that for all $1 \le i, j \le n$, we have $\|\hat{J}_{ij} - J^\star_{ij}\| \le \tilde{\alpha}/2$, where $\|.\|$ is the norm of interest and $\tilde{\alpha}$ is the required precision.

2. *Structure reconstruction:* Compute an estimate $\hat{E}$ of E such that the probability of perfect reconstruction satisfies $p[\hat{E} = E] \ge 1 - \delta$ where δ defines the confidence.

Most existing methods in the literature (Vuffray et al., 2016; 2019; Klivans & Meka, 2017) use parameter estimation to perform structure reconstruction. It is evident that whenever the parameters can be estimated with precision $\tilde{\alpha} = \alpha$, the structure estimated by the thresholding procedure given by

$$\hat{E}(\alpha) = \{1 \le (i,j) \le n \mid |\hat{J}_{ij}| \ge \alpha/2\}, \quad (3)$$

results in a perfect reconstruction with high probability. An information theoretic lower bound for the dynamic model selection problem from a single trajectory was derived in

(Bresler et al., 2017) and is given by

$$m \geq \frac{e^{2\beta d/3}}{32 d\alpha e^{d+3\beta+6}} n \log n \tag{4}$$

In comparison, the information-theoretic sample complexity for learning from i.i.d. samples (Santhanam & Wainwright, 2012) scales as $\exp(\beta d)$. It still remains unclear if either of the information theoretic lower bounds are tight. Current evidence or constructive proofs show a scaling of $\exp(4\beta d)$ for the i.i.d. case (Lokhov et al., 2018) and $\exp(20\beta d)$ for the general dynamic case (Bresler et al., 2017).

3. Learning Ising models from dynamics

3.1. Learning algorithms

We now describe how to adapt RISE and RPLE into computationally efficient estimators for learning Ising models from Glauber dynamics. Both algorithms minimize a convex loss function that relies on the properties of the conditional distributions rather than the full probability distribution. These methods reconstruct the neighborhoods of each node independently and are, therefore, fully parallelizable. Moreover, the minimization procedure can be implemented in $\tilde{O}(n^2)$ using entropic gradient descent for both RISE (Vuffray et al., 2019) and RPLE (see (Klivans & Meka, 2017) for a related stochastic first-order method with multiplicative updates).

Unlike the i.i.d. sample setting, the Glauber dynamics naturally takes the form of a local neighborhood update rule conditioned on the event that the spin i is updated, see Eq. (2). Following the construction of the RISE estimator in (Vuffray et al., 2016), we define the Dynamic Interaction Screening Objective for each node $i \in V$ as being the inverse of the exponent of the conditional distribution,

$$\textbf{D-ISO: } \mathcal{S}_m(\underline{J}_i, H_i) = \frac{1}{m_i}\sum_{t=1}^{m} \exp\left[-\sigma_i^{t+1}\left(\sum_{j\neq i} J_{ij}\sigma_j^t + H_i\right)\right]\delta_{i,I^{t+1}}, \tag{5}$$

where $\underline{J}_i := \{J_{ij} \mid j \neq i\} \in \mathbb{R}^{n-1}$ denotes the vector of pairwise interactions around a node i, and $m_i = \sum_{t=1}^{m} \delta_{i,I^{t+1}}$. The Kronecker delta $\delta_{i,I^{t+1}}$ in Eq. (5) is used to keep samples for which updates happened at i.

The estimators' objectives have been stated considering the samples come from the T-regime. Similar expressions are stated for the M-regime in Appendix S1.

We call the corresponding estimator which uses D-ISO as the *Dynamic Regularized Interaction Screening Estimator (D-RISE)* and is defined in the spirit of (Vuffray et al., 2016) as the following convex program,

$$\textbf{D-RISE: } (\hat{\underline{J}}_i, \hat{H}_i) = \underset{(\underline{J}_i, H_i)}{\operatorname{argmin}} \mathcal{S}_m(\underline{J}_i, H_i) + \lambda||\underline{J}_i||_1, \tag{6}$$

where the ℓ_1-regularization promotes sparsity and λ is a tunable parameter controlling the amount of sparsity enforced.

The pseudo-likelihood based estimator can be understood as the (negative) conditional likelihood (Ravikumar et al., 2010) of an update at node i and takes the following form in the case of Glauber dynamics,

$$\textbf{D-PL: } \mathcal{L}_m(\underline{J}_i, H_i) = -\frac{1}{m_i}\sum_{t=1}^{m} \ln\left[1 + \sigma_i^{t+1}\tanh\left(\sum_{j\neq i} J_{ij}\sigma_j^t + H_i\right)\right]\delta_{i,I^{t+1}}. \tag{7}$$

Analogous to D-RISE, the *Dynamic Regularized Pseudo-Likelihood Estimator (D-RPLE)* takes the form of an ℓ_1-regularized convex program,

$$\textbf{D-RPLE: } (\hat{\underline{J}}_i, \hat{H}_i) = \underset{(\underline{J}_i, H_i)}{\operatorname{argmin}} \mathcal{L}_m(\underline{J}_i, H_i) + \lambda||\underline{J}_i||_1 \tag{8}$$

The performance of the learning algorithms depends on the regularization parameter λ. Setting it too high encourages the interaction parameters to drop out and setting it too low can make the estimation sensitive to noise. Following theoretical considerations explained further, a good choice for successfully reconstructing the local neighborhood of i with probability $1-\delta'$ (where $\delta' = \delta/n$ and $1-\delta$ is the success of the whole graph reconstruction) is to set $\lambda = c_\lambda\sqrt{\log(n^2/\delta')/m_i}$ where the intensity of the penalty increases in a logarithmic fashion with the size of the system n and decreases with the number of spin updates observed m_i. The parameter $c_\lambda > 0$ is a numerical constant independent of the problem parameters such as m_i and n.

One of our main contributions is the following theorem which quantifies the sample complexity required for structure learning from Glauber dynamics in the M-regime.

Theorem 1 (M-regime: Structure Learning of Ising Model Dynamics)**.** *Let $\{\underline{\sigma}^{0(t)}, \underline{\sigma}^{1(t)}, I^{1(t)}\}_{t\in[m]}$ be m samples of spin configurations and corresponding node identities drawn through Glauber dynamics (Eq. 2), and define $m_i = \sum_{t=1}^{m}\delta_{i,I^{1(t)}}$ as the number of updates per spin i. Consider M-regime on an Ising model with maximum degree d, maximum coupling intensity β, minimum coupling intensity α, and for simplicity assume $H_i^\star = 0\ \forall i$. Then for any $\delta > 0$, the following estimators with penalty parameter of form $\lambda \propto \sqrt{\log(3n^3/\delta)/m_i}$ reconstruct the edge-set perfectly with probability $p(\hat{E}(\lambda,\alpha) = E) \geq 1-\delta$ if the number of samples satisfies*

i) D-RPLE: $m_i \geq C_d \max(1, \alpha^{-2})\exp(4\beta d)\ln(3n^3/\delta)$,

ii) D-RISE: $m_i \geq C'_d \max(1, \alpha^{-2})\exp(2\beta d)\ln(3n^3/\delta)$,

where C_d and C'_d depend only polynomially on d.

A more precise statement and proof of Theorem 1 is given in Appendix S1. Notice that given the choice of the initial distribution $p(\underline{\sigma}_0)$ to be the uniform distribution, the total number of samples m required to get the number of samples m_i that satisfies Theorem 1 is $m = O(nm_i)$. Notice that unlike the i.i.d. case where the entire sample may be distinct, in the dynamic case only a single spin is updated while the values of other variables are kept fixed; hence, perhaps a more natural quantity for comparison with the i.i.d. case is the number of updates per spin m_i instead of m.

We expect the worst-case scalings of learning from dynamics in the T-regime to be similar to the i.i.d. setting as it includes the fully mixed setting as a particular case. The main contribution of Theorem 1 lies in that it establishes with certainty that the scaling of D-RISE in the M-regime is strictly better than in the i.i.d. setting where it was found experimentally to be at least $\exp(4\beta d)$ in the worst case (Lokhov et al., 2018). It is also interesting to compare the above results to the theoretical analysis of the i.i.d. setting for which RPLE and RISE have scalings upper-bounded by $\exp(8\beta d)$ and $\exp(6\beta d)$ respectively (Lokhov et al., 2018). However, these theoretical upper-bounds tend to be loose and this motivates us to quantify the scalings achieved in practice in the dynamic case experimentally.

3.2. Empirical β scaling of the sample complexity

Our main goal is to assess the empirical sample complexity of our learning algorithms in both the T-regime and M-regime. In particular, we want to extract the exponential scaling of the sample complexity for successful structure reconstruction with respect to β, the magnitude of the largest coupling. The sample complexity of D-RPLE and D-RISE are tested over Ising models of different topologies and interaction strengths. We do not include a comparison to the algorithm of (Bresler et al., 2017) due to its high computational complexity, and to heuristic mean-field methods (Roudi & Hertz, 2011; Mézard & Sakellariou, 2011; Zeng et al., 2011; Zhang, 2012; Bachschmid-Romano & Opper, 2015) since most of them are derived for asymteric kinetic Ising model, and are not guaranteed to reconstruct arbitrary strongly interacting models; instead, we focus on studying the performance of two exact methods that can be efficiently implemented through convex optimization.

We denote the minimal number of samples required for perfect structure reconstruction with probability $1 - \delta \geq 0.95$ (for $\delta = 0.05$) as $m^\star$. Our experimental protocol to find $m^\star$ (sample complexity of the structure learning problem) is similar to the one from Supplementary material of (Lokhov et al., 2018). For each graphical model topology and coupling values, we determined $m^\star$ by finding the minimal value of m samples required to successively reconstruct the structure 45 times in a row from 45 independent sets of m samples, which guarantees a 90% confidence for $\delta = 0.05$.

We reconstruct the topology by first solving the optimization problems in Eq. (6) for D-RISE and in Eq. (8) for D-RPLE with appropriate ℓ_1 regularization to obtain estimates of $(\hat{\underline{J}}_i, \hat{H}_i)$ at each node $i \in V$. We create a consensus of the estimated couplings by averaging the reconstruction from two nodes i.e. $\hat{J}_{ij}^{\text{avg}} = (\hat{J}_{ij} + \hat{J}_{ji})/2$. The set of pairwise interactions $\hat{\underline{J}}^{\text{avg}}$ which are higher than $\alpha/2$ are defined as the estimate of the edge set $\hat{E}$. The structure is declared to be successfully reconstructed when the edge set is perfectly recovered $\hat{E} = E$.

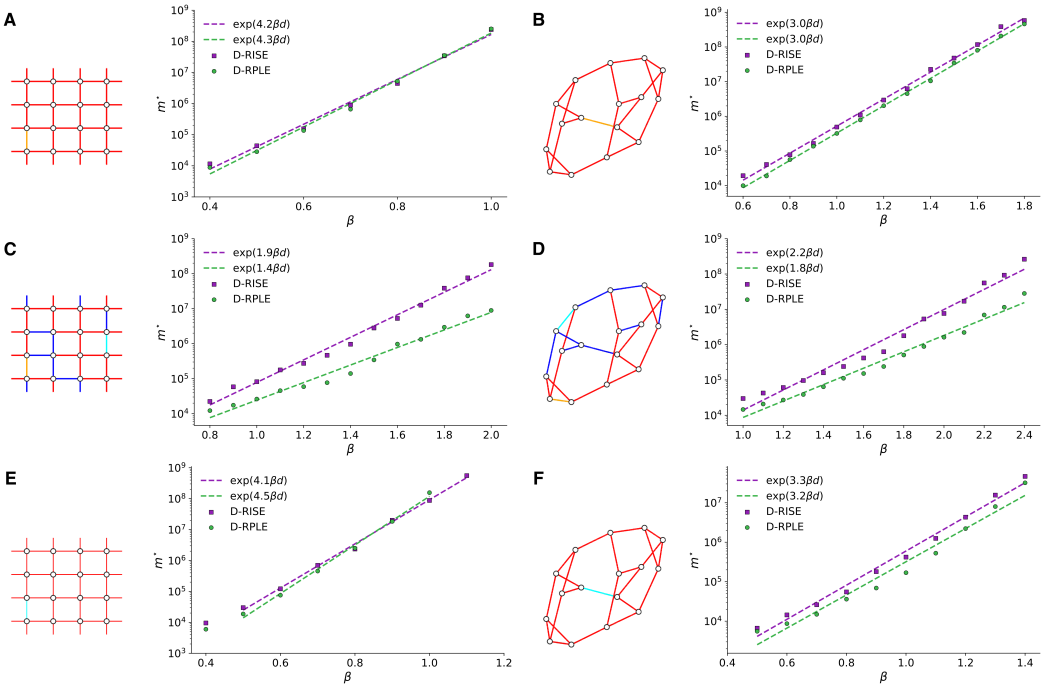
We perform an extensive set of numerical experiments to empirically obtain $m^\star$ for a variety of graphs in both the T-regime and the M-regime. We consider two different topologies: the periodic two-dimensional lattice and the random 3-regular graph. Each of these two topologies are subdivided into three categories according to the signs of the interaction couplings. This includes the ferromagnetic case with positive couplings, the spin glass case with couplings taking random signs and the ferromagnetic case with a weak anti-ferromagnetic impurity (i.e., weak negative coupling). For each of these six cases, all the couplings' magnitudes are set to $|J_{ij}^\star| = \beta$ with exception of one or two couplings which are set to $|J_{ij}^\star| = \alpha$. In our experiments, we fixed the value of $\alpha = 0.4$ and varied β from α to a value ranging from 1 to 4 depending on the model. All models have their magnetic fields at each node $H_i^\star$ set to zero. These cases are identical to the test cases used in (Lokhov et al., 2018) in the i.i.d. learning setting which enables us to draw a comparison between the dynamic and i.i.d. settings.

In deciding the ℓ_1-regularization to be used, optimal values of c_λ which are unknown apriori were determined through detailed numerical simulations on different Ising model topologies as described in Appendix S3. The determined optimal values of c_λ are summarized in Table 1 on lattices and random regular (RR) graphs for the two different regimes.

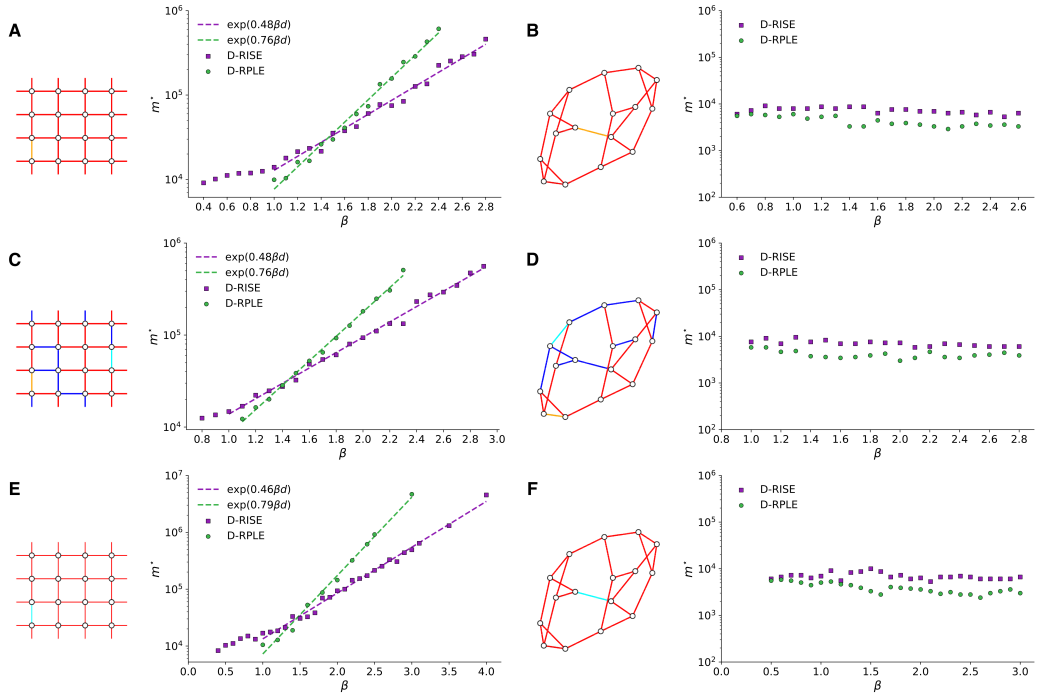
	T-regime		**M-regime**	
	D-RISE	**D-RPLE**	**D-RISE**	**D-RPLE**
Lattices	0.1	0.05	0.1	0.05
RR Graphs	0.45	0.1	0.7	0.3

Table 1: Optimal values of c_λ for D-RISE and D-RPLE.

Our sample complexity results for the T-regime and M-regime are shown in Figure 1a and Figure 1b respectively. In the T-regime, the worst scalings of sample complexity are observed for both D-RISE and D-RPLE in lattices with purely ferromagnetic interactions (Fig. 1aA) and those with a weak anti-ferromagnetic interaction (Fig. 1aE). The worst cases are $\exp(4.2\beta d)$ and $\exp(4.5\beta d)$ for D-RISE and D-RPLE respectively. The scaling of sample complexity results are similar to the i.i.d. setting for the ferromagnetic models on lattices and random regular graphs. However,



(a) Scaling of $m^\star$ with β in T-regime.



(b) Scaling of $m^\star$ with β in M-regime.

Figure 1: We assess the performance of D-RISE and D-RPLE in reconstructing Ising models of size $n = 16$ from samples generated from Glauber dynamics. The different Ising model topologies with their corresponding pictorial representations on the left-hand side of each plot are: (A) ferromagnetic model on a periodic lattice, (B) spin glass model on a periodic lattice, and (C) ferromagnetic model on a periodic lattice with weak antiferromagnetic impurity. Edges in the pictorial representations of the models are colored red (β), orange (α), cyan ($-\alpha$) and blue ($-\beta$). Value of $\alpha = 0.4$ for all the graphs.

compared to the i.i.d. case, we obtain a 35% reduction in the β scaling for D-RISE for the spin glass model on a lattice and around 27% reduction for the spin glass model on a random regular graph. For the systems studied here, the Glauber dynamics mixes rapidly in the case of ferromagnetic models as the number of variables is small compared to the number of samples required to learn the structure. Therefore the dynamics in the T-regime quickly produces samples that behave like i.i.d. samples in these cases and we see similar scalings. This mixing may not be as rapid in the case of spin glass models, yielding us savings in number of samples when learning from Glauber dynamics.

The picture drastically changes as we move to the M-regime for which the scaling results are shown in Fig. 1b. Among the numerical examples that we consider, the worst-case scaling for D-RISE and D-RPLE observed are $\exp(0.48\beta d)$ and $\exp(0.79\beta d)$ respectively, with the scalings being very similar for all the lattices. Compared to the T-regime there is a reduction in the β scaling between 80% and 90%, translating to a reduction in the sample requirement of orders of magnitude. Interestingly, we observe that a constant number of samples independent of β is required for learning the random regular graphs. Thus, it is possible to beat exponential scalings for special topologies in the case of M-regime which would not be possible in the i.i.d. setting. The details of this behavior is described in Appendix S4.

The fundamental difference between the two regimes is that the Glauber dynamics comes effectively from two different initial distributions. In the M-regime, the samples are produced from a one step Glauber dynamics initialized with a uniform distribution. In the T-regime, however, the samples are *effectively* produced from a one step Glauber dynamics that is initialized from a distribution which is close to the equilibrium one, as the actual dynamics mixes more rapidly. This shows that the dynamical samples acquired far from the equilibrium carry more information about the structure of graphical models. A natural question to ask is then how to empirically find such an initial distribution for the Glauber dynamics that improves the sample complexity. We propose a possible solution to this issue by introducing an active learner in Section 4.2.

4. Applications

In this section, we illustrate how D-RISE/D-RPLE can be applied to a real world system and extended to improve their performance. In Section 4.1, we consider a multi-neuron spike trains' data set to learn the structure of a network of neurons. This can be used to understand the dynamics of the network and how it implements computations. In Section 4.2, we highlight an approach to modify the initial distribution to the M-regime to improve sample complexity.

4.1. Learning Ising models from neural data

Due to the high dimensionality of the space of spike patterns and lack of enough data to build an exact statistical description, it has become popular to use parametric models such as Ising models (Schneidman et al., 2006b). In the corresponding Ising model (Rieke et al., 1999), the spin σ_i of neuron/node i can be interpreted as spiking/firing ($\sigma_i = 1$) or not ($\sigma_i = -1$). Studies on using Ising models for spike trains include application to larger populations of neurons (Cocco et al., 2009; Nirenberg & Victor, 2007), conditions under which Ising models are good approximations (Roudi et al., 2009a; Tkacik et al., 2009), development of faster learning methods (Broderick et al., 2007) and comparisons of different learning methods (Roudi et al., 2009b). Most of previous studies consider the i.i.d. setting. However, (Hertz et al., 2011) showed that respecting correlations in time and the dynamics can lead to better Ising model fits to the data. This motivates us to investigate the underlying Ising model for spike trains considering Glauber dynamics.

We consider the dataset from (Prentice et al., 2016) containing spike trains from 152 salamander retinal ganglion cells in response to a non-repeated natural movie stimulus, of which we select spike trains over $n = 42$ neurons over 24s for our application. To use D-RISE/D-RPLE, the dataset is first converted into a sequence of 1.2×10^5 spin configurations, a segment of which is shown as a spike raster in Figure 2. Time series of spin configurations along with updated node identities are then extracted from this sequence and 3.2×10^4 samples corresponding to the M-regime with an unknown initial distribution are obtained. Details of this procedure is given in Appendix S5.

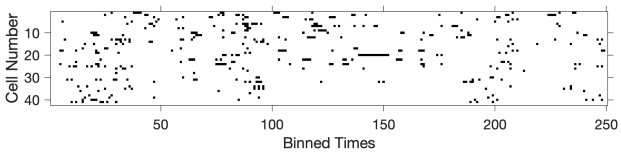
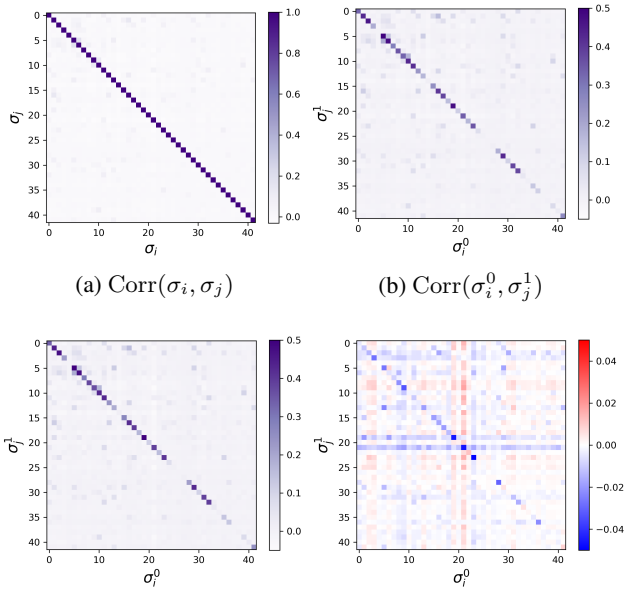


Figure 2: Spike raster from the first 5 sec of the data over 42 neurons. Each column indicates the spiking pattern of the neurons over a 20 ms time bin.

We discuss the statistics of Ising model parameters learned using D-RISE on this set of samples in Appendix S5, where we also compare the recovered parameters with those obtained by RISE assuming the samples are i.i.d. Correlations computed from data assuming the samples are i.i.d. (Figure 3a) and that respecting time (Figure 3b) are very different. This difference strengthens the importance of respecting dynamics when learning an effective Ising model if one would like to capture time correlations present in the data. The correlation matrix predicted using the model learned with D-RISE is presented in Figure 3c, and is within $\sim 10\%$ of that computed from data under the Frobenius

norm (see Figure 3d), indicating a good model fit that respects the time correlations present in the data. Details of correlation computation can be found in the Appendix S5.

As we have control in such biological systems through external stimuli, learning an effective Ising model could be accelerated using an active learner which we discuss next.



(a) $\mathrm{Corr}(\sigma_i, \sigma_j)$ (b) $\mathrm{Corr}(\sigma_i^0, \sigma_j^1)$

(c) Predicted $\mathrm{Corr}(\sigma_i^0, \sigma_j^1)$ (d) Difference in predicted and data correlations

Figure 3: Correlation matrices are computed from data in (a) assuming the spin configurations are i.i.d., and (b) respecting time ordering (see definition in Appendix S5). We show the predicted correlation matrix using D-RISE estimates in (c) and its difference from that computed from data in (d).

4.2. Active learning in M-regime

Motivated by the previous section, we develop an active learning algorithm that optimizes on the fly over the initial distribution in the M-regime. In the M-regime, the initial spin configuration $\underline{\sigma}^0$ can be viewed as a query to the Glauber dynamics which returns the output of $(\underline{\sigma}^1, I^1)$. Clearly, the observations generated have a dependence on the query distribution $p_{\underline{\sigma}^0}$ from which the queries $\underline{\sigma}^0$ are simulated. If no prior information about the parameter set $(\underline{J}, \underline{H})$ is known, then a suitable choice of $p_{\underline{\sigma}^0}$ is the uniform distribution over $\{-1, +1\}^n$ as we had chosen for our numerical experiments in the previous section.

In *active learning*, the learner has the ability to select queries during model learning that would be most informative. Here, we consider a *mini-batch adaptive active learning* (Wei et al., 2015) scheme where in each round a mini-batch of initial spin configurations $\underline{\sigma}^0$ are selected to be queried and then the samples obtained are combined with those from before to produce estimates of $(\hat{\underline{J}}, \hat{\underline{H}})$. These estimates are then used to determine the next mini-batch of queries. To select these queries we use the informative measure of entropy as in uncertainty sampling. An alternate criteria that can be used is Fisher information but the computational effort of the resulting query optimization typically grows exponentially with n (Sourati et al., 2017a). In uncertainty sampling (Settles, 2009), one query $\underline{\sigma}^0$ is chosen at a time

$$\hat{\underline{\sigma}}^0 = \underset{\underline{\sigma}^0 \in \{-1,+1\}^n}{\operatorname{argmax}} \; S(\underline{\sigma}^1 | \underline{\sigma}^0; \hat{\underline{J}}, \hat{\underline{H}}) \tag{9}$$

where S is the entropy measure of the probability $p(\underline{\sigma}^1 | \underline{\sigma}^0; \hat{\underline{J}}, \hat{\underline{H}})$ based on current parameter estimates. The entropy in the case of Glauber dynamics is

$$S(\underline{\sigma}^1 | \underline{\sigma}^0) = \sum_{k \in [n]} \log(2\cosh(A_k)) - A_k \tanh(A_k) \tag{10}$$

where $A_k = \sum_{l \in \partial k} J_{kl} \sigma_l^0 + H_k$. Here, we issue mini-batches of queries sampled from distribution q that is proportional to the the entropy S. Our algorithm is given in Algorithm 1.

Algorithm 1 Active Learning of Glauber Dynamics

Input: Initial set of samples $X^{(0)}$, number of mini-batches i_{max}, size of mini-batch m_b
  $(\hat{\underline{J}}, \hat{\underline{H}}) \leftarrow$ D-RISE$(X^{(0)})$
  for $i = 1 : i_{max}$ **do**
    Compute entropy $S(\underline{\sigma}^1 | \underline{\sigma}^0; \hat{\underline{J}}, \hat{\underline{H}}) \; \forall \underline{\sigma}^0$
    Set $q(\underline{\sigma}^0) \propto S(\underline{\sigma}^1 | \underline{\sigma}^0; \hat{\underline{J}}, \hat{\underline{H}})$
    Modify distribution: $q = \mu q + (1 - \mu) p_U$
    Sample m_b queries from $\{-1, +1\}^n$ w.p. q
    Obtain corresponding samples X_b by running Glauber dynamics in M-regime
    Set $X^{(i)} = X^{(i-1)} \bigcup X_b$
    $(\hat{\underline{J}}, \hat{\underline{H}}) \leftarrow$ D-RISE$(X^{(i)})$
  end for
Output: $(\hat{\underline{J}}, \hat{\underline{H}})$

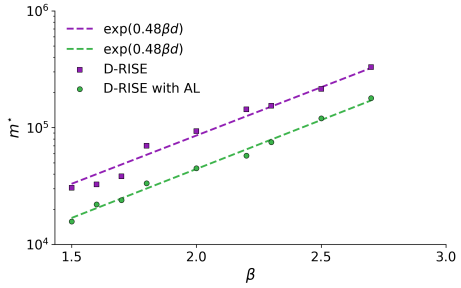


Figure 4: Scaling of $m^\star$ with β in M-regime. Performance comparison of D-RISE with active learning against a vanilla D-RISE in reconstructing a ferromagnetic model with weak anti-ferromagnetic impurity of size $n = 16$ (see Fig. 1bE).

Note that we slightly modify the query distribution q for each mini-batch by mixing it with the uniform distribution p_U. The mixing coefficient $0 \leq \mu \leq 1$ typically depends on the number of samples so far. We set it to $\mu = 1 - 1/|X^{(i)}|^{1/6}$ which is often used for such active learning algorithms (Sourati et al., 2017b; Chaudhuri et al., 2015).

To determine the sample complexity, the minimal number of samples required for successful structure reconstruction $m^\star$ is determined as described in Sec. 3.2. However, here each trial corresponds to an independent active learning run. In each trial for a given value of m, we consider the size of initial set of samples to be $\lfloor m/3 \rfloor$ and size of batches m_b such that there are a total of 15 mini-batches.

In Figure 4, we compare the sample complexity of D-RISE with and without active learning (AL) on the challenging case of a ferromagnetic periodic lattice with a weak antiferromagnetic impurity (Fig. 1bE). We consider values of β between 1.5 and 2.7. While the scaling of sample complexity with β remains unchanged (within experimental error), there is about 47% reduction in the number of samples required for successful graph reconstruction when using AL.

5. Conclusions and future work

In this paper, we theoretically and empirically establish a fundamental difference between learning graphical models in the traditional framework of i.i.d. samples and samples obtained from out of equilibrium dynamics. We show that in the latter understudied setting, there is considerable improvement in sample complexity which has both theoretical and practical consequences. In future work, it would be interesting to further investigate general Markov Random Fields and other Markov chain dynamics.

Code, data availability, and supplementary material

The code for the learning algorithms, active learner and data in this work is available on GitHub[1]. The supplementary material can be found at the following link [2].

Acknowledgements

A.D. was partially supported by the Applied Machine Learning Fellowship at LANL where most of the work was carried out. A.Y.L., M.V., and S.M. acknowledge support from the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project numbers 20190059DR, 20200121ER, and 20210078DR.

[1] https://github.com/lanl-ansi/learning-ising-dynamics

[2] https://arxiv.org/abs/2104.00995

References

Acharya, J., Bhattacharyya, A., and Kamath, P. Improved bounds for universal one-bit compressive sensing. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 2353–2357, 2017. doi: 10.1109/ISIT.2017.8006950.

Bachschmid-Romano, L. and Opper, M. Learning of couplings for random asymmetric kinetic ising models revisited: random correlation matrices and learning curves. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(9):P09016, 2015.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Ben-Tal, A., Margalit, T., and Nemirovski, A. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12 (1):79–108, 2001.

Berry, M. J., Warland, D. K., and Meister, M. The structure and precision of retinal spike trains. *Proceedings of the National Academy of Sciences*, 94(10):5411–5416, 1997.

Besag, J. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195, 1975.

Biegler, L. T. and Zavala, V. M. Large-scale nonlinear programming using ipopt: An integrating framework for enterprise-wide dynamic optimization. *Computers & Chemical Engineering*, 33(3):575–582, 2009.

Boufounos, P. T. and Baraniuk, R. G. 1-bit compressive sensing. In *2008 42nd Annual Conference on Information Sciences and Systems*, pp. 16–21, 2008. doi: 10.1109/CISS.2008.4558487.

Bresler, G. Efficiently learning Ising models on arbitrary graphs. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pp. 771–782. ACM, 2015.

Bresler, G., Gamarnik, D., and Shah, D. Learning graphical models from the glauber dynamics. *IEEE Transactions on Information Theory*, 64(6):4072–4080, 2017.

Broderick, T., Dudik, M., Tkacik, G., Schapire, R. E., and Bialek, W. Faster solutions of the inverse pairwise ising problem. *arXiv preprint arXiv:0712.2437*, 2007.

Buczak, A. L. and Guven, E. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys Tutorials*, 18 (2):1153–1176, 2016.

Chaudhuri, K., Kakade, S. M., Netrapalli, P., and Sanghavi, S. Convergence rates of active learning for maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, pp. 1090–1098, 2015.

Chaves, R., Majenz, C., and Gross, D. Information–theoretic implications of quantum causal structures. *Nature communications*, 6:5766, 2015.

Chow, C. and Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, May 1968. ISSN 1557-9654. doi: 10.1109/TIT.1968.1054142.

Cocco, S., Leibler, S., and Monasson, R. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proceedings of the National Academy of Sciences*, 106(33):14058–14062, 2009.

Constantinou, A. C., Fenton, N., Marsh, W., and Radlinski, L. From complex questionnaire and interviewing data to intelligent bayesian network models for medical decision support. *Artificial intelligence in medicine*, 67:75–93, 2016.

Decelle, A. and Zhang, P. Inference of the sparse kinetic ising model using the decimation method. *Physical Review E*, 91(5):052136, 2015.

Decelle, A., Ricci-Tersenghi, F., and Zhang, P. Data quality for the inverse lsing problem. *Journal of Physics A: Mathematical and Theoretical*, 49(38):384001, 2016.

Eagle, N., Pentland, A. S., and Lazer, D. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36): 15274–15278, 2009. doi: 10.1073/pnas.0900282106.

Glauber, R. J. Time-dependent statistics of the ising model. *Journal of mathematical physics*, 4(2):294–307, 1963.

Gotovos, A., Hassani, H., and Krause, A. Sampling from probabilistic submodular models. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1945–1953. Curran Associates, Inc., 2015.

Hamilton, L., Koehler, F., and Moitra, A. Information theoretic properties of markov random fields, and their algorithmic applications. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2463–2472. Curran Associates, Inc., 2017.

He, M. and Zhang, J. A dependency graph approach for fault detection and localization towards secure smart grid. *IEEE Transactions on Smart Grid*, 2(2):342–351, June 2011. ISSN 1949-3053. doi: 10.1109/TSG.2011.2129544.

Hertz, J., Roudi, Y., and Tyrcha, J. Ising models for inferring network structure from spike data. *arXiv preprint arXiv:1106.1752*, 2011.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. A bayesian networks approach for predicting protein-protein interactions from genomic data. 302(5644):449–453.

Kahn, J., Komlós, J., and Szemerédi, E. On the probability that a random$\pm$1-matrix is singular. *Journal of the American Mathematical Society*, 8(1):223–240, 1995.

Klivans, A. and Meka, R. Learning graphical models using multiplicative weights. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 343–354, Oct 2017.

Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Lokhov, A. Y., Vuffray, M., Misra, S., and Chertkov, M. Optimal structure and parameter learning of ising models. *Science advances*, 4(3):e1700791, 2018.

Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., and Stolovitzky, G. Wisdom of crowds for robust gene network inference. *Nat Meth*, 9(8):796–804, Aug 2012. ISSN 1548-7091. doi: 10.1038/nmeth.2016.

Marre, O., El Boustani, S., Frégnac, Y., and Destexhe, A. Prediction of spatiotemporal patterns of neural activity from pairwise correlations. *Physical review letters*, 102 (13):138101, 2009.

Martinelli, F. and Olivieri, E. Approach to equilibrium of glauber dynamics in the one phase region. *Communications in Mathematical Physics*, 161(3):447–486, 1994.

Mézard, M. and Sakellariou, J. Exact mean-field inference in asymmetric kinetic ising systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(07):L07001, 2011.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011. doi: 10.1073/pnas.1111471108.

Negahban, S., Yu, B., Wainwright, M. J., and Ravikumar, P. K. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in neural information processing systems*, pp. 1348–1356. Citeseer, 2009.

Nirenberg, S. H. and Victor, J. D. Analyzing the activity of large populations of neurons: how tractable is the problem? *Current opinion in neurobiology*, 17(4):397–400, 2007.

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., and Simoncelli, E. P. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

Prentice, J. S., Marre, O., Ioffe, M. L., Loback, A. R., Tkačik, G., and Berry, M. J. Error-robust modes of the retinal population code. *PLoS computational biology*, 12 (11):e1005148, 2016.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D., et al. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Rieke, F., Warland, D., Van Steveninck, R. D. R., Bialek, W. S., et al. *Spikes: exploring the neural code*, volume 7. MIT press Cambridge, 1999.

Roth, S. and Black, M. J. Fields of experts: a framework for learning image priors. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pp. 860–867 vol. 2, June 2005. doi: 10.1109/CVPR.2005.160.

Roudi, Y. and Hertz, J. Mean field theory for nonequilibrium network reconstruction. *Physical review letters*, 106(4): 048702, 2011.

Roudi, Y., Nirenberg, S., and Latham, P. E. Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't. *PLoS Comput Biol*, 5(5):e1000380, 2009a.

Roudi, Y., Tyrcha, J., and Hertz, J. Ising model for neural data: model quality and approximate methods for extracting functional connectivity. *Physical Review E*, 79(5): 051915, 2009b.

Santhanam, N. P. and Wainwright, M. J. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.

Schneidman, E., Berry, M. J., Segev, R., and Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440 (7087):1007–1012, Apr 2006a. ISSN 0028-0836. doi: 10.1038/nature04701.

Schneidman, E., Berry, M. J., Segev, R., and Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087): 1007–1012, 2006b.

Settles, B. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Sly, A. and Sun, N. The computational hardness of counting in two-spin models on d-regular graphs. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pp. 361–369. IEEE, 2012.

Sourati, J., Akcakaya, M., Erdogmus, D., Leen, T. K., and Dy, J. G. A probabilistic active learning algorithm based on fisher information ratio. *IEEE transactions on pattern analysis and machine intelligence*, 40(8):2023–2029, 2017a.

Sourati, J., Akcakaya, M., Leen, T. K., Erdogmus, D., and Dy, J. G. Asymptotic analysis of objectives based on fisher information in active learning. *The Journal of Machine Learning Research*, 18(1):1123–1163, 2017b.

Tkacik, G., Schneidman, E., Berry II, M. J., and Bialek, W. Spin glass models for a network of real neurons. *arXiv preprint arXiv:0912.5409*, 2009.

Tyrcha, J., Roudi, Y., Marsili, M., and Hertz, J. The effect of nonstationarity on models inferred from neural data. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03005, 2013.

Vuffray, M., Misra, S., Lokhov, A., and Chertkov, M. Interaction screening: Efficient and sample-optimal learning of ising models. In *Advances in Neural Information Processing Systems*, pp. 2595–2603, 2016.

Vuffray, M., Misra, S., and Lokhov, A. Y. Efficient learning of discrete graphical models. *arXiv preprint arXiv:1902.00600*, 2019.

Wang, C., Komodakis, N., and Paragios, N. Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding*, 117(11):1610 – 1627, 2013. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2013.07.004.

Wei, K., Iyer, R., and Bilmes, J. Submodularity in data subset selection and active learning. In *International Conference on Machine Learning*, pp. 1954–1963, 2015.

Zeng, H.-L., Aurell, E., Alava, M., and Mahmoudi, H. Network inference using asynchronously updated kinetic ising model. *Physical Review E*, 83(4):041135, 2011.

Zeng, H.-L., Alava, M., Aurell, E., Hertz, J., and Roudi, Y. Maximum likelihood reconstruction for ising models with asynchronous updates. *Physical review letters*, 110 (21):210601, 2013.

Zhang, P. Inference of kinetic ising model on sparse graphs. *Journal of Statistical Physics*, 148(3):502–512, 2012.